Comparative analysis of prosodic characteristics using WaveNet embeddings

Suni, Antti

ISCA
2019-09-15

Suni , A , Wlodarczak , M , Vainio , M & Simko , J 2019 , Comparative analysis of prosodic characteristics using WaveNet embeddings . in G Kubin , T Hain , B Schuller , D El Zarka & P Hodl (eds) , 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019) : Crossroads of Speech and Language . Interspeech , ISCA , Baixas , pp. 2538-2542 , Annual Conference of the International Speech Communication Association , Graz , Austria , 15/09/2019 . https://doi.org/10.21437/Interspeech.2019-2373

http://hdl.handle.net/10138/307676
https://doi.org/10.21437/Interspeech.2019-2373

unspecified
publishedVersion

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Comparative analysis of prosodic characteristics using WaveNet embeddings

Antti Suni[1], *Marcin Włodarczak*[2], *Martti Vainio*[1], *Juraj Šimko*[1]

[1]University of Helsinki, Finland
[2]Stockholm University, Sweden

{juraj.simko,antti.suni,martti.vainio}@helsinki.fi
wlodarczak@ling.su.se

Abstract

We present a methodology for assessing similarities and differences between language varieties and dialects in terms of prosodic characteristics. A multi-speaker, multi-dialect WaveNet network is trained on low sample-rate signal retaining only prosodic characteristics of the original speech. The network is conditioned on labels related to speakers' region or dialect. The resulting conditioning embeddings are subsequently used as a multi-dimensional characteristics of different language varieties, with results consistent with dialectological studies. The method and results are illustrated on a Swedia 2000 corpus of Swedish dialectal variation.

Index Terms: language comparison, prosodic typology, WaveNet, embeddings

1. Introduction

Natural language processing approaches have recently been used to learn distributed language representations that reflect typological relationships among languages. Language embeddings trained within text-based multi-language machine translation systems have been shown to contain typological information that can be used to derive language family trees closely resembling the trees previously obtained by meticulous linguistic analysis [1, 2, 3].

In addition to linguistic features, such as syntax, morphology and lexicon, the languages show systematic similarities and differences also in their prosodic characteristics. This is true in particular for dialects and varieties of the same language. In this paper we present and evaluate a methodology for comparative typological prosodic analysis of language varieties based on state-of-the-art deep network language modeling approach (WaveNet). The presented method provides an alternative to several signal-based approaches to prosodic typology published in the past [4, 5, 6, 7].

WaveNet is an artificial neural network speech synthesis platform trained to generate raw audio samples in an autoregressive probabilistic fashion. Given a suitable corpus, the system can learn voice characteristics of multiple speakers in parallel, encoded in the form of learnt embeddings used as a conditioning of the network [8]. We explore the possibility of learning distributed representations of prosodic characteristics of language varieties in a similar manner and evaluate the extent to which the resulting embeddings contain meaningful typological information.

We use a WaveNet synthesis system to train distributed dialect representations in a form of embeddings of global conditioning of the network. The network is conditioned (in addition to standard conditioning by previous waveform samples) through two serially implemented embeddings. The first conditioning is applied by feeding the ID of a language variety – in the form of one-hot encoded vector – through an embedding; this *target embedding* is assumed to capture typological relationships between the varieties. The second embedding is designed to capture and normalize the sources of variation that are necessarily present in the data but are not deemed relevant to the typological task: speaker's sex and age as well as lexical properties of the material. As the aim is typological anaysis based on prosody only, the network is trained on a low sample-rate signal retaining only *prosodic* characteristics of the original speech material, namely its f_0 and energy. These signals used are sinusoidal waveforms with instantaneous frequency equal to the fundamental frequency of the original speech samples (two-syllabic words from a dialectal database of Swedish) modulated by the energy envelope from the original signal. The waveforms are downsampled to 800 Hz sampling rate, enough to capture f_0 within a range present in the data.

In standard WaveNet synthesis, modelling prosody is challenging due to long time scales involved. Instead, separately predicted f_0 trajectories are usually provided as local conditioning features to the model. Apart from only keeping prosodically relevant signal properties, the low sample rate signal used here naturally increases the time step between two subsequent samples, and in effect considerably expands the temporal extent of perceptive field, i.e., the samples effectively conditioning generation of the next sample in the feedforward convolutional WaveNet architecture. The receptive field of the tested networks is 1024 samples corresponding to 1.28 s at 800 Hz. As a consequence, the trained language model takes into account temporal context long enough to encompass prosodically relevant phenomena in the given corpus.

In summary, we can model the long-range phenomena related to suprasegmental prosody, while still operating with one-dimensional time-domain signal where the dependencies between energy, pitch and duration are intact.

2. WaveNet-based prosodic clustering

2.1. Network design

A globally conditioned generative WaveNet deep neural network was trained to generate downsampled speech signal waveforms. A TensorFlow implementation of network architecture described in detail in [8] was used in this work.

Briefly, the WaveNet architecture learns to generate probability distributions of quantized sample values of a raw audio signal by processing a stretch of the previous samples through a stack of dilated convolutional layers. At generation time, the predicted sample (selected based on the predicted distribution) is directly fed back as part of the input of the network in an auto-regressive fashion. The stacked dilated convolutional layers increase the size of the receptive filed for the prediction (i.e, the length of the previous portion of the signal that conditions

http://dx.doi.org/10.21437/Interspeech.2019-2373

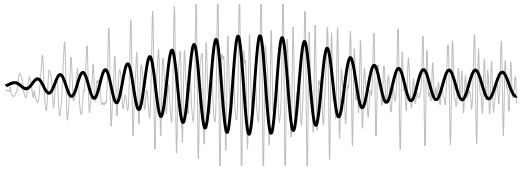

Figure 1: *A part of the original waveform (in grey) and the corresponding sinusoidal signal (in black).*

sample generation). Gradually increasing dilation of each subsequent layer also provides a sort of parallel hierarchical analysis with more dilated layers capturing progressively longer-term dependencies in the signal.

In addition to conditioning by the previous signal, the WaveNet architecture allows implementation of global conditioning. The relevant characteristics of a given signal (language, speaker's id, speakers sex, ...) are fed as an additional input to each dilated convolutional network trough embeddings trained alongside the other network components. The network can thus learn to produce audio signal with the required characteristics. In essence, when trained, each embedding layer maps a discrete set of relevant parameters, e.g., a one-hot encoded set of language labels, to a real number valued vector directly used to condition each convolutional layer. Presumably, this mapping captures mutual relationships within the labeled material: intuitively, the more similar the waveforms to be generated the more similar the conditioning vectors.

2.2. Prosody Signal

The aim of this work is to evaluate relationships concerning prosodic characteristics of speech material. That is, we are trying to extract statistical similarities and differences in signal attributes such as f_0 and intensity patterns, while disregarding segmental properties of the signal.

In order to preserve the required signal characteristics and suppress the other properties, the speech signal used for training the networks was pre-processed as follows:

First, f_0 contours (in Hz) were extracted from the speech material using YAAPT fundamental frequency tracking procedure [9]; subsequently, the unvoiced intervals were interpolated (and extrapolated). Sinusoidal signal with the instantaneous frequency given by the f_0 contours (i.e., the same as the fundamental frequency of the original signal) and the same energy envelope as the original signal was generated:

$$s(t) = e(t)\sin\big(2\pi \int_0^t f_0(\tau)d\tau\big),$$

where $e(t)$ is the energy envelope of the original signal (see Fig. 1). The signal $s(t)$ was then downsampled to 800 Hz and scaled to the range -1 and 1. A μ-law companding transformation was applied to reduce the dynamic range and the waveforms were quantized to 256 possible values.

A more straightforward approach, namely utilizing lowpass-filtered and downsampled original speech signals was also tested, with the benefit of not requiring explicit f_0 extraction. But similarly to what has been observed in other dialect corpora [10], the models trained with such raw signals were better in differentiating the recording conditions[1] than the dialectal variation of the sites.

2.3. Embeddings

Although the implemented network is generative and thus learns to produce the downsampled signals, we are primarily interested in the embeddings implementing global conditioning. We assume that the embeddings quantify the prosodically relevant differences and similarities among different categories used as global conditionings of the network.

Two parallel global conditioning embeddings were incorporated in the WaveNet implementation used here.

The first embedding layer, which we refer to as *target embedding*, maps a one-hot encoded category label of interest (recording location, or a combination of geographical and phonological information, see below) to a multi-dimensional real-valued conditioning vector. These conditioning vectors corresponding to individual labels will be used as representations of the relevant learnt prosodic characteristics pertaining to given categories of interest.

The downsampling process described above keeps intact some properties of the signal that might be interfering with the aims of typological analysis. In particular, the processed signals are not normalized with respect to gender-based pitch differences and differences in linguistic material. To counteract this source of variability, we use a second embedding layer, called here *normalization embedding*. This layer simply conditions the network through a learnt embedding of labels combining sex and age group of the speakers and the uttered words (see below).

3. Evaluation

3.1. Material

The WaveNet based prosodic clustering approach as well as the speaker normalization technique described above was evaluated on Swedia 2000, a corpus of Swedish dialects spoken around Sweden and Swedish-speaking areas in Finland, collected in the years 1998–2003 [11].

A subset of SweDia 2000 containing isolated words was used for training. This part contains 253,725 recordings (individual words, read as sequences of several repetitions of each word), recorded in 104 locations with multiple speakers (12 on average) for each location. Only two-syllabic words (14 different Swedish words) and the first two repetitions in a sequence were used for training and validation, resulting in 25,491 tokens in total (20,477 for training and 5,044 for validation).

Each recording is labeled by a recording location and the location's geographical region: (mainland) Finland, Åland, and (from south northwards in mainland Sweden) Götaland, Svealand, Norrland. The recordings also contain speakers' age and sex information (four categories: younger or older female or younger or older male).

Every recording location was assigned to one of the five Swedish tonal dialect types. Briefly, regional variation in the realization of the two Swedish word accent types (acute and grave), associated with the stressed syllable, has been traditionally described in terms of two binary features: (i) the number of pitch peaks in the grave accent (type 1 and type 2 dialects), and (ii) the timing of the pitch peak in the acute accent (type A and type B dialects). In addition to the resulting dialectal categories (1A, 1B, 2A, 2B), type 0 dialects do not use contrastive word accents [12]. The tonal dialect labels in SweDia were added by [13] using a method based on the classification of neighbour-

[1] Specifically, the Swedia 2000 corpus recordings from the northern (Norrland, Östrabotten) and southern (Götaland) regions contain distinguishable mains hum of 50 Hz, absent in the recordings from central (Svealand) region.

ing sites in [12]. We excluded all uncertain cases for which this method failed to produce a single dialectal category.

In an earlier study [7], we have used the Swedia 2000 word lists for automatic classification of the Swedish tonal dialects using a wavelet-based hierarchical prosodic analysis. Our results indicated that the traditional dialectal categories are mediated by geographical proximity, i.e., neighbouring sites belonging to different dialectal categories are prosodically similar.

3.2. Network parameters and training

In order to evaluate whether the models capture typologically relevant information, we carried out two separate clustering experiments: one using embeddings for individual recording sites, and another one using embeddings for the geographical regions (Norrland, Svealand, Götaland, Åland and Finland) and tonal dialects (0, 1A, 1B, 2A, 2B). In a more general view, the first model is largely unsupervised, with no dialectal knowledge applied. Should meaningful results be achieved, such model could be applied on languages where information of dialects and their geographical distribution is more limited. In contrast, the second experiment demonstrates how the proposed framework can be applied to complement and test existing typological descriptions.

For both experiments, we built a WaveNet model with 18 dilated convolutional layers arranged in two stacks, yielding a receptive field of 1024 samples or 1.28 seconds, covering the full length of the short utterances in our training data. The numbers of residual and skip channels were set to 32 and 128, respectively. Both normalization and target embedding dimensions were set to 32 in Experiment 1 (Section 3.3.1) whereas the target embedding was reduced to 16 dimensions in Experiment 2 (Section 3.3.2). The trained embedding values were normalized by subtracting the means from the columns corresponding to the embedding dimensions.

Models were trained using Adam optimizer (learning rate of 0.001), with batch size of approximately 10 words (8 seconds). Training was stopped when the loss on validation set failed to decrease for 20 consecutive epochs. Training was repeated several times, and the resulting embeddings were checked. All the trained embeddings yielded qualitatively similar results.

3.3. Results

3.3.1. Site-based embeddings

Fig. 2 visualizes the normalization embeddings of labels combining speakers' sex and age, and the uttered word, transformed using principal component analysis. Along the first principal component, the embedding clearly reflects separation by sex, as well as age, in particular for females. Presumably, the separation reflects mean f_0 of the speakers' groups, including lower average pitch for older females compared to younger ones and somewhat higher for older than for younger males [14]. The second principal component then captures linguistic content of the utterances with, as indicated on a few examples, similar topology of the word embeddings for all sex-age groups.

The left panel in Fig. 3 shows a k-means-based clustering (n=5) of the target embedding vectors corresponding to individual recording sites. The sites cluster in a geographically relatively homogeneous and meaningful way. The cluster depicted in green contains predominantly Finnish sites. In Sweden, the blue cluster contains mostly northern recording sites and largely overlaps with the Norrland region. The cluster shown in magenta contains mostly sites from central Sweden around the capital, roughly corresponding to Svealand region. The red and orange clusters cover most sites in the southern Götaland region, with the "orange" cluster containing the southernmost sites (Skåne) and the sites from the Gotland island. The clustering broadly corresponds to the traditional geographical distribution of Swedish dialects [12].

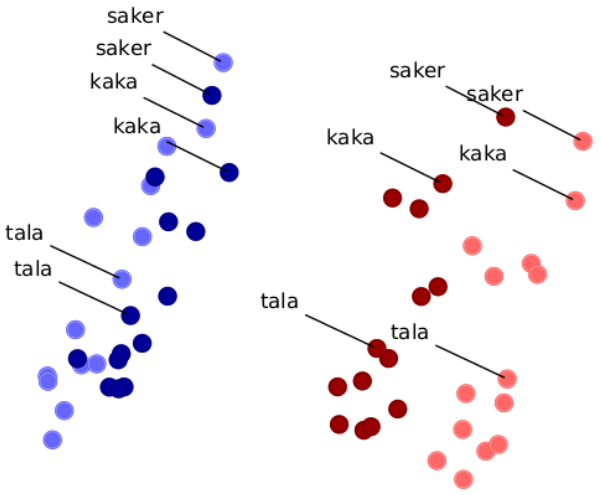

Figure 2: *First two principal components of the normalization embedding. (blue: male, red: female; light: young, dark: old)*

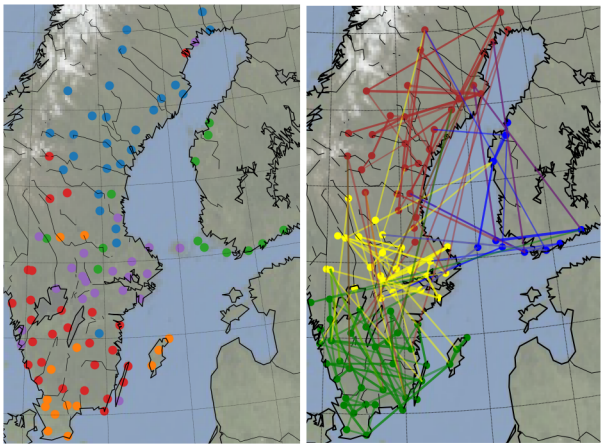

Figure 3: *Left: k-means-based clustering (n=5) of the embedding vectors for individual recording sites. Right: recording locations connected with their prosodic nearest neighbors; the edges are colored by geographical region (green: Götaland, yellow: Svealand, red: Norrland, blue: Finland).*

A strong effect of geographical proximity is also demonstrated by the results in the right panel of Fig. 3, where each recording site was connected to its two nearest neighbours in terms of distance between the embedding vectors. It is clear that the edges predominantly join locations within geographical regions with connections crossing region boundaries (e.g. from Svealand to Norrland) being exceedingly rare.

To quantify the degree to which the embedding vectors capture geographical distribution of the sites, we compared the mutual Euclidean distances among the embedding vectors (*embeddist*) with geographical distances of the corresponding sites

(*geodist*). A linear regression with vector distance as a dependent and geographical distance as an independent variable shows significantly positive slope ($p < 0.001$), and yields significant adjusted R^2 value of 0.054 (corresponding to significantly positive correlation of 0.23 between these distance variables).

To test the degree to which embedding distance depends on dialectal characteristics of recording sites, we created two factor variables based on tonal dialect distinctions: variable *numtype* capturing the main tonal dialect type (0, 1 vs 2) of the pair of sites, and *ABtype* containing the subtypes (0, A or B) of the site pairs. Each of these variables has six levels corresponding to all possible type comnbinations (0-0, 0-1, 0-2, 1-1, 1-2, 2-2 for *numtype*; 0-0, 0-A, 0-B, A-A, A-B, B-B for *ABtype*).

Table 1: *Adjusted R^2 values of linear models predicting embedding vector distance (embeddist), from geographical distance (geodist), main tonal dialect types (numtype) and the ABtypes of location pairs.*

Model	Adjusted R^2
embeddist $\sim$ *geodist*	0.055
embeddist $\sim$ *geodist * ABtype*	0.116
embeddist $\sim$ *geodist * numtype*	0.146
embeddist $\sim$ *geodist * numtype * ABtype*	0.154

Including these two variables as independent variables alongside geographical distance in the linear regression model with embedding vector distance as a dependent variable increases the quality of fit of the model in the way summarized in Table 1. In all cases reported in the table, the more complex model explained significantly more variance in the data than the simpler model, suggesting that the properties of the tonal dialects of the recording sites are indeed reflected in the corresponding embedding vectors.

3.3.2. Tonal dialect and region based embeddings

To evaluate the relative influence of geography and tonal dialect explicitly, another model was trained with the target embeddings based on conditioning by a geographical region and tonal dialect type to each recording. Five geographically distinct regions (Norrland, Svealand, Götaland, Åland and Finland were used as a proxy of geographical locations of individual sites. The region information was combined with one of the five possible tonal dialect type (0, 1A, 1B, 2A, 2B). This lead to 11 region-dialect combinations, as not all dialectal types are spoken in every region.

Fig. 4 shows a dendrogram obtained by hierarchical clustering based on Euclidean distances among the resulting 11 embedding vectors (R fuction `hclust` was used). The primary split runs along major dialect type, perfectly separating types 0, 1 and 2. Within these dialectal types, however, the geographical influences dominate the influence of the subtype (A vs B).

4. Discussion

The evaluation shows that typologically meaningful clustering based on purely prosodic information can be obtained by the proposed methodology. Also, we have shown that the second embedding containing information about speakers' sex and age as well as lexical particulars of the word tokens can be used to effectively disentangle these influences from the typological

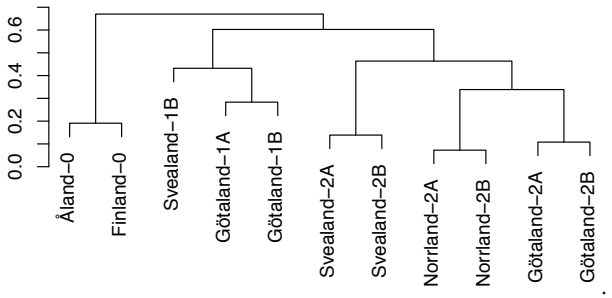

Figure 4: *Dendrogram of region and dialect type combinations.*

characteristics of the material.

The clustering of geographically subdivided tonal accent types largely agree with the previous analysis of the same corpus presented in [7]. Namely, the embedding topology reflects the primary division of the speech tokens based on major dialect types, but the division based on timing of pitch peaks is obscured by geographical proximity of recording sites. We believe that this interaction between tonal dialects and geography is a genuine typological phenomenon. Admittedly, however, it is possible that the method fails to robustly distinguish between pitch contours differing in timing because there is no precise information on syllable structure available in the dataset (other than the energy envelope).

The meaningful typological information have been extracted from the material well suited to this type of analysis, namely read word lists, where word accent realization is known to vary between dialects. It can be assumed that the results might somewhat differ for other types of material (e.g., spontaneous speech, longer read stories, etc.) even for the same language. The current methodology provides a means of analysing any given speech material by constructing a conditional probabilistic language model trained as a part of a generative synthesis system.

As mentioned in the introduction, pre-processing of the data (replacing the original waveforms by downsampled sinusoids with the same fundamental frequency and energy envelope, see Section 2.2) provided two benefits: (1) a guarantee that the typological analysis was done based on purely prosodic features, and (2) the expansion of the temporal context used to construct the probabilistic language model. In addition, this step may bring an additional advantage, namely removing potential spectral "noise" that might be present in analysed data due to challenging recording conditions, (moderate amount of) cross-talk, etc. This might prove to be advantageous when applying the proposed method to speech material often used for typological analysis, such as field recordings.

5. Acknowledgements

The work was funded in part by the Academy of Finland grant no. 1293348 *Digital Language Typology* to the first author, and by a Stockholm University *docentstipentium* to the second author. We would like to thank Vassilis Tsiaras from the University of Crete for providing us with the WaveNet implementation that was adapted for this work, and Johan Frid from Lund University for sharing his dialect labels for the Swedia 2000 corpus.

6. References

[1] R. Östling and J. Tiedemann, "Continuous multilinguality with language vectors," in *15th Conference of the European Chapter*

of the Association for Computational Linguistics, 2017, pp. 644–649.

[2] J. Bjerva and I. Augenstein, "Tracking typological traits of uralic languages in distributed language representations," in *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, 2018, pp. 76–86.

[3] J. Bjerva, R. Östling, M. H. Veiga, J. Tiedemann, and I. Augenstein, "What do language representations really represent?" *arXiv preprint arXiv:1901.02646*, 2019.

[4] F. Cummins, F. Gers, and J. Schmidhuber, "Automatic discrimination among languages based on prosody alone," Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland, Tech. Rep., 1999.

[5] J. Šimko, A. Suni, K. Hiovain, and M. Vainio, "Comparing languages using hierarchical prosodic analysis," in *Proc. Interspeech 2017*, Stockholm, Sweden, 2017.

[6] K. Hiovain, A. S. Suni, J. Simko, and M. Vainio, "Mapping areal variation and majority language influence in North Sámi using hierarchical prosodic analysis," in *Proc. Speech Prosody 2018*, Poznań, Poland, 2018.

[7] M. Włodarczak, J. Šimko, A. Suni, and M. Vainio, "Classification of Swedish dialects using a hierarchical prosodic analysis," in *Proc. Speech Prosody 2018*, Poznań, Poland, 2018.

[8] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[9] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[10] H. Boril, A. Sangwan, and J. H. L. Hansen, "Arabic dialect identification - 'is the secret in the silence?' and other observations," in *INTERSPEECH*, 2012.

[11] A. Eriksson, "SweDia 2000: A Swedish dialect database," in *Babylonic Confusion Resolved: Proceedings of the Nordic Symposium on the Comparison of Spoken Languages*, ser. Copenhagen Working Papers in LSP, P. J. Henrichsen, Ed., no. 1, 2004, pp. 33–48.

[12] E. Gårding, *The Scandinavian word accents*. Lund: Gleerup, 1977.

[13] J. Frid, "Lexical and acoustic modelling of Swedish prosody," Ph.D. dissertation, Lund University, Lund, 2003.

[14] M. L. Stoicheff, "Speaking fundamental frequency characteristics of nonsmoking female adults," *Journal of Speech, Language, and Hearing Research*, vol. 24, no. 3, pp. 437–441, 1981.